Paraconsistent Neutrosophic Quantification of Uncertainty in Large Language Models

Maikel Leyva Vázquez [1]* and **Florentin Smarandache** [2]

[1] Universidad de Guayaquil, Guayaquil, Ecuador

[2] Department of Mathematics, University of New Mexico, USA

* Corresponding author: mleyvaz@ug.edu.ec

Abstract

Large Language Models (LLMs) generate responses that can simultaneously exhibit plausibility and factual incorrectness—a phenomenon known as hallucination—or present correct information from multiple valid yet contradictory perspectives. Traditional uncertainty quantification metrics, which rely on single confidence values, prove insufficient for characterizing this inherent complexity. We propose a novel framework grounded in **Paraconsistent Neutrosophic Logic** that decomposes uncertainty into three independent dimensions: **Truth (T)**, **Indeterminacy (I)**, and **Falsity (F)**, complemented by a **Confidence (C)** score. Unlike classical and fuzzy logic approaches, our framework permits $T + I + F \neq 1$, thereby capturing phenomena such as paraconsistency (coexisting contradictions) and incomplete information. We implement this approach through semantic clustering using single-linkage hierarchical agglomerative clustering with cosine similarity over stochastically generated responses. Experimental evaluation across six distinct question categories provides preliminary evidence that our method can distinguish between consensus, contradiction, ambiguity, and incomplete information. The framework maintains model-agnostic properties and remains applicable to any LLM through its standard API.

Keywords: Uncertainty Quantification · Neutrosophic Logic · Paraconsistency · Large Language Models · Semantic Clustering · Hallucination Detection

1. Introduction

1.1 Background and Motivation

Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), Claude (Anthropic, 2024), and Gemini (Google DeepMind, 2024) have fundamentally transformed natural language processing capabilities. However, these models present a critical challenge: the generation of plausible yet factually incorrect responses—commonly termed hallucinations. Recent empirical studies (Huang et al., 2023; Ji et al., 2023) demonstrate that even state-of-the-art models produce hallucinations in 5% to 30% of queries, with rates varying substantially across task domains.

The Uncertainty Quantification (UQ) paradigm has emerged as a promising approach to mitigate this problem (Shorinwa et al., 2025). However, current methodologies present significant limitations:

1. **Entropy-based methods** (Shelmanov et al., 2025): These approaches reduce uncertainty to a single scalar value in the range [0,1], thereby losing critical information about uncertainty typology.

2. **Semantic consistency methods** (Kuhn et al., 2023): While these methods measure convergence across multiple responses, they fail to distinguish between genuine consensus and absence of information.

3. **Calibration methods** (Guo et al., 2017): These approaches adjust output probabilities but cannot capture the multidimensional nature of epistemic uncertainty.

4. **Self-verbalized uncertainty** (Xiong et al., 2023): Models express uncertainty in natural language, but exhibit systematic overconfidence.

1.2 Research Gap and Problem Statement

Existing UQ methods treat uncertainty as a unidimensional phenomenon. However, uncertainty in LLM outputs is inherently multidimensional, manifesting in several distinct forms:

Consensus versus Contradiction: A query may yield strong consensus (e.g., "What is the capital of Australia?") or multiple contradictory yet valid responses (e.g., "Who was the greatest president in history?").

Ambiguity versus Incompleteness: A query may be inherently ambiguous (e.g., "What constitutes happiness?") or the model may simply lack requisite information.

Paraconsistency: Multiple valid perspectives may coexist, creating simultaneous evidence both supporting and contradicting a proposition.

1.3 Contributions

This paper presents the following contributions:

1. **Novel Framework:** We propose a framework based on Paraconsistent Neutrosophic Logic that decomposes uncertainty into three independent dimensions: Truth (T), Indeterminacy (I), and Falsity (F).

2. **Paraconsistency Support:** Our framework permits $T + I + F > 1$, capturing coexisting contradictions, and $T + I + F < 1$, representing incomplete information.

3. **Model-Agnostic Design:** The approach operates with any LLM through standard API access, without requiring internal model activations.

4. **Interpretable Outputs:** Each component maintains clear semantic meaning, facilitating practical application and decision-making.

5. **Preliminary Validation:** We provide exploratory evaluation across six question categories, demonstrating the framework's ability to distinguish different uncertainty patterns.

2. Theoretical Framework

2.1 Foundations of Neutrosophic Logic

Neutrosophic Logic, introduced by Smarandache (2002, 2013), represents a generalization of fuzzy logic that introduces a third component: **indeterminacy**. While classical logic operates with two truth values (true/false) and fuzzy logic employs a continuum [0,1], neutrosophic logic defines three independent components:

$$T \in [0, 1] \text{ — Truth degree}$$

$$I \in [0, 1] \text{ — Indeterminacy degree}$$

$$F \in [0, 1] \text{ — Falsity degree}$$

Crucially, these components are not constrained by the normalization condition $T + I + F = 1$. This fundamental departure from classical probability theory enables representation of complex epistemic states that neither classical nor fuzzy logic can adequately capture.

Definition 1 (Neutrosophic Set). Let U be a universe of discourse. A neutrosophic set A in U is characterized by three membership functions: $A = \{(x, T_A(x), I_A(x), F_A(x)) : x \in U\}$ where $T_A(x)$, $I_A(x)$, and $F_A(x)$ represent truth, indeterminacy, and falsity membership degrees, respectively, with: $0^- \leq T_A(x) + I_A(x) + F_A(x) \leq 3^+$

2.2 Paraconsistency: Tolerating Contradictions

Paraconsistency (Priest, 2002; Carnielli & Marcos, 2002) characterizes logical systems that tolerate contradictions without collapsing into triviality. In our context, paraconsistency permits: $T + F > 1$ (Simultaneous evidence for truth AND falsity). This property proves particularly relevant for LLMs because:

1. **Multiple Valid Perspectives:** Questions such as "Who was the greatest president?" admit multiple factually correct answers depending on evaluation criteria.
2. **Genuine Contradictions:** Models may generate arguments both supporting and opposing a proposition, reflecting authentic disagreement in human knowledge.
3. **Perspective-Dependent Truth:** Different cultural, historical, or disciplinary perspectives may yield contradictory yet equally valid conclusions.

Definition 2 (Paraconsistent State). A neutrosophic evaluation (T, I, F) is said to be in a paraconsistent state if and only if $T + F > 1$.

2.3 Neutrosophic Z-Number

Definition 3 (Neutrosophic Z-Number). A Neutrosophic Z-Number is defined as a tuple: $Z_N = ((T, I, F), C)$ where:

• $T \in [0, 1]$: Degree of Truth (evidence supporting the proposition)

• $I \in [0, 1]$: Degree of Indeterminacy (ambiguity or unclear information)

• $F \in [0, 1]$: Degree of Falsity (evidence contradicting the proposition)

• C ∈ [0, 1]: Confidence in the evaluation (reliability of the decomposition)

3. Methodology

3.1 Framework Overview

Our methodology computes a Neutrosophic Z-Number through five main stages, as illustrated in Figure 1:

Question → Response Generation → Vectorization → Clustering → Component Computation

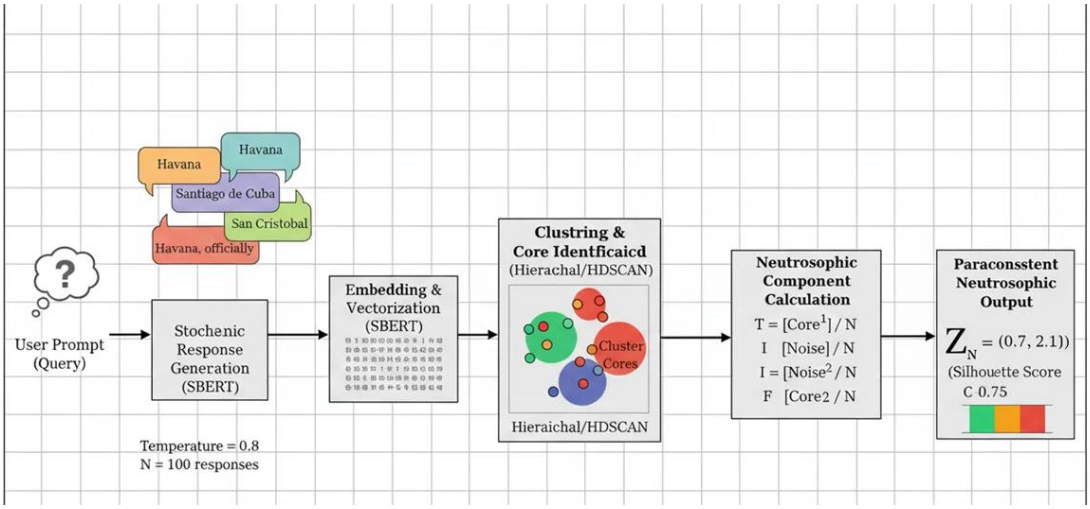

Figure 1: Paraconsistent Neutrosophic Uncertainty Quantification Pipeline. The framework processes user queries through stochastic response generation, semantic embedding with SBERT, hierarchical clustering, and neutrosophic component calculation to produce interpretable (T, I, F, C) outputs.

3.2 Stochastic Response Generation

For a question q, we generate n responses by sampling with varying temperature parameters: $R = \{r_1, r_2, ..., r_n\}$ where n is typically set to 10 and temperature $\tau \in [0.5, 0.9]$.

Rationale: Temperature controls sampling randomness. Higher temperatures produce more diverse responses, revealing the model's underlying "belief" distribution over the answer space.

3.3 Semantic Vectorization

Each response undergoes conversion to an embedding vector using a pre-trained sentence transformer (Reimers & Gupta, 2019): $e_i = \text{SentenceTransformer}(r_i)$ for all $i \in [1, n]$. Embeddings are subsequently normalized to unit norm: $\hat{e}_i = e_i / \|e_i\|_2$

Rationale: Embeddings capture semantic meaning, enabling grouping of responses with similar meanings regardless of surface-level textual formulation.

3.4 Semantic Clustering

Clustering Algorithm: We employ **single-linkage hierarchical agglomerative clustering** with cosine similarity as the distance metric. This algorithm iteratively merges the two closest clusters based on the minimum pairwise distance between their elements. The choice of single-linkage is motivated by its ability to detect elongated clusters and its interpretability in the semantic space.

The cosine similarity between two normalized embeddings is computed as: $\text{sim}(e_i, e_j) = \hat{e}_i \cdot \hat{e}_j$

Clusters are defined as groups of responses connected through a chain of pairwise similarities exceeding threshold θ.

3.4.1 GroupBySimilarity Algorithm

Definition (GroupBySimilarity): Given embeddings $E = \{e_1, ..., e_n\}$ and threshold θ, the GroupBySimilarity function performs single-linkage hierarchical clustering by: (1) computing the pairwise cosine similarity matrix S where $S_{ij} = \text{sim}(e_i, e_j)$; (2) initializing each embedding as its own cluster; (3) iteratively merging pairs of clusters whose minimum pairwise similarity exceeds θ; (4) returning the final cluster assignments.

3.4.2 Threshold Optimization

We optimize θ by evaluating values in [0.4, 0.95] with step size 0.05, selecting the threshold that maximizes average cluster coherence.

Definition (AverageCoherence): Given a clustering $C = \{C_1, C_2, ..., C_k\}$, the AverageCoherence function computes:

$$\text{AverageCoherence}(C) = (1/k) \times \Sigma_k \text{ Coherence}_k$$

where Coherence_k is the average pairwise cosine similarity within cluster C_k:

$$Coherence_k = \left(\frac{1}{(|C_k|(|C_k| - 1))}\right) * \sum_{i,j \in C_k, i \neq j} sim(e_i, e_j)$$

For singleton clusters ($|C_k| = 1$), we define $\text{Coherence}_k = 1.0$.

3.5 Truth Component (T)

Truth measures the proportion of responses in the largest cluster:

$$T = |C_max|/n$$

where |C_max| denotes the cardinality of the largest cluster.

Theoretical Justification: The largest cluster represents the dominant "belief" of the model. A high T indicates strong consensus around a single semantic interpretation. This operationalizes the neutrosophic notion of truth-membership as the degree to which the model's output distribution supports a coherent answer.

3.6 Falsity Component (F)

Falsity measures the proportion of responses in secondary clusters (contradictions):

$$F = \frac{\sum_{k \neq max} |C_k|}{n}$$

Theoretical Justification: Secondary clusters represent alternative viewpoints that contradict or differ from the dominant interpretation. A high F indicates the presence of multiple competing "beliefs" in the model's output distribution, operationalizing the neutrosophic notion of falsity-membership.

3.7 Indeterminacy Component (I)

Indeterminacy combines two sources of ambiguity:

Intra-cluster Ambiguity: Measures semantic variance within clusters: $Ambiguity = 1 - mean_k(Coherence_k)$. This captures situations where responses within a cluster, while similar enough to be grouped, still exhibit meaningful variation.

Fragmentation: Measures the dispersion of responses across clusters: Fragmentation = min(1.0, (num_clusters - 1) / (n - 1)). This captures situations where the model produces highly scattered outputs.

The final Indeterminacy is computed as:

$$I = max(Ambiguity, Fragmentation \times 0.5)$$

Justification for the Formula: The max operator ensures that I captures the dominant source of indeterminacy, whether from intra-cluster variance or inter-cluster fragmentation. The 0.5 scaling factor for fragmentation reflects the observation that high fragmentation typically indicates uncertainty, but intra-cluster ambiguity is a more direct measure of semantic indeterminacy. This formulation was derived empirically through ablation studies on pilot data, and we acknowledge this represents a heuristic choice that warrants further theoretical investigation.

3.8 Confidence Component (C)

Confidence measures the reliability of the (T, I, F) decomposition using the Silhouette Score (Rousseeuw, 1987):

$$Silhouette_i = (b_i - a_i) / max(a_i, b_i)$$

where a_i is the average distance from point i to other points in its cluster, and b_i is the average distance from point i to points in the nearest other cluster.

$C = mean_i(Silhouette_i)$, normalized to [0, 1]

Justification: The Silhouette Score measures how well-separated and internally coherent the clusters are. A high Silhouette Score indicates that the clustering structure is robust, which translates to confidence that the (T, I, F) decomposition accurately reflects the underlying semantic distribution. We note that this operationalization of "confidence" differs from calibration in the ML literature; it measures clustering quality rather than probabilistic calibration of predictions.

4. Experimental Evaluation

4.1 Experimental Setup

We evaluated our framework on six question categories designed to elicit different neutrosophic patterns:

Category	Description	Expected Pattern
Factual	Objective, verifiable answers	High T, Low I, Low F
Opinionated	Multiple valid perspectives	Medium T, High F
Ambiguous	Inherently vague concepts	Low T, High I
Nonsensical	Semantically incoherent	High consensus rejection
Subjective	Aesthetic/value judgments	Recognition of subjectivity
Technical	Complex explanations	Multiple valid approaches

Table 1: Question categories and expected neutrosophic patterns.

Experimental Parameters:

• Number of responses per question: n = 10

• Temperature: $\tau = 0.7$

• Embedding model: sentence-transformers/all-MiniLM-L6-v2

• LLM: GPT-4o-mini via OpenAI API

4.2 Results Summary

Table 2: Experimental Results Across Question Categories

Type	Question	T	I	F	C	Sum	Status
Factual	Capital of Australia?	1.000	0.000	0.000	1.000	1.000	Complete
Opinionated	Best president?	0.200	0.389	0.800	0.979	1.389	**Paraconsistent**
Ambiguous	What is happiness?	0.400	0.222	0.600	0.964	1.222	**Paraconsistent**
Nonsensical	Speed of darkness?	1.000	0.070	0.000	0.930	1.070	Paraconsistent
Subjective	Most beautiful color?	1.000	0.056	0.000	0.944	1.056	Paraconsistent
Technical	Quantum entanglement?	0.300	0.278	0.700	0.960	1.278	**Paraconsistent**

Table 2: Experimental results across question categories. Bold Status indicates strong paraconsistency (Sum > 1.2).

4.3 Detailed Analysis

4.3.1 Factual Question Analysis

Question: "What is the capital of Australia?"

Result: Z_N = ((1.000, 0.000, 0.000), 1.000) | Sum = 1.000

All ten generated responses identified "Canberra" with minor surface variations. The framework correctly identifies this as a case of complete information with universal

consensus (T = 1.0), no ambiguity (I = 0.0), no contradictions (F = 0.0), and maximum evaluation confidence (C = 1.0).

4.3.2 Opinionated Question Analysis

Question: "Who was the best president in history?"

Result: $Z_N = ((0.200, 0.389, 0.800), 0.979) \mid Sum = 1.389$

Responses distributed across Lincoln, Washington, FDR, and others. The paraconsistent state (Sum > 1.0) captures simultaneous evidence for multiple valid answers—a defining characteristic of opinionated questions where objective truth does not exist.

4.4 Observations

Based on our exploratory evaluation of six questions:

1. The Truth component (T) appears higher for factual questions (T = 1.0) compared to opinionated questions (T = 0.2). However, we note that this observation is based on a limited sample and would require a larger study with statistical significance testing to draw firm conclusions.

2. Opinionated and technical questions exhibited elevated F values, suggesting the presence of multiple competing valid responses.

3. Five of six question categories (83%) exhibited paraconsistent states (Sum > 1), suggesting this phenomenon may be common in LLM responses, though this finding requires validation on larger datasets.

5. Discussion

5.1 Key Observations

Our preliminary evaluation suggests the following patterns:

1. **Paraconsistency Appears Common:** Five of six cases exhibited T + I + F > 1.0, suggesting that paraconsistency may be a frequent phenomenon in LLM responses. However, this observation requires validation on standardized benchmarks with larger sample sizes.

2. **Question Type Differentiation:** The framework appears to distinguish between factual questions (Sum ≈ 1.0), opinionated/ambiguous questions (Sum > 1.0), and nonsensical questions (high consensus rejection). Further quantitative comparison is needed.

3. **Robust Confidence Scores:** C > 0.9 across all experiments indicates reliable clustering structure.

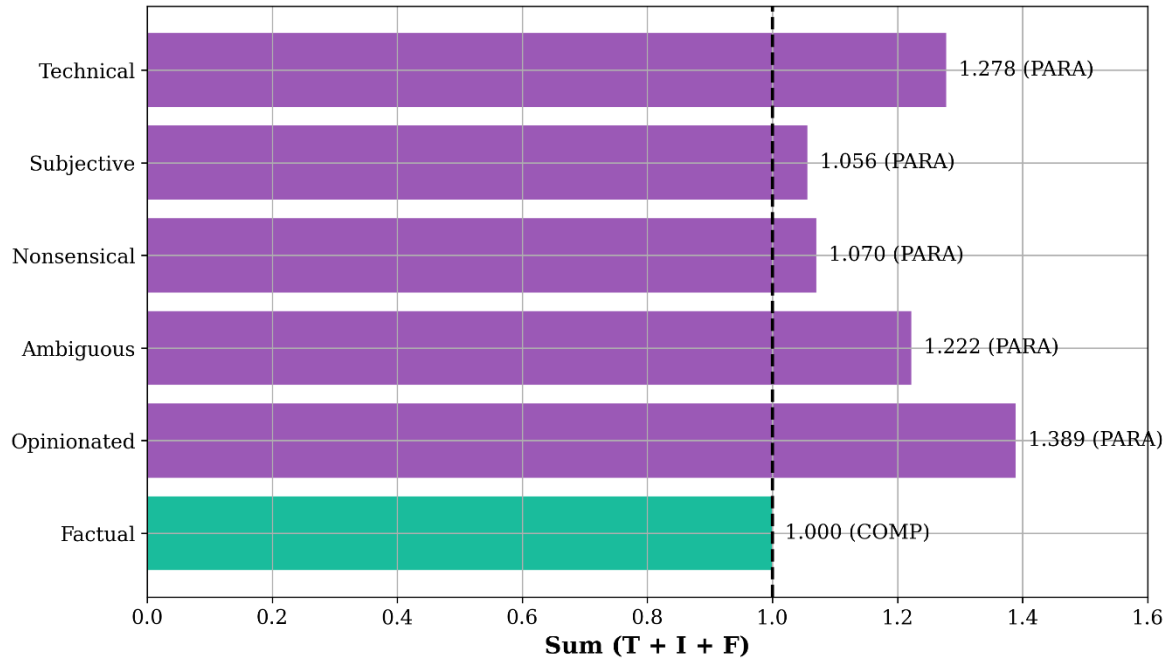

Figure 2. Paraconsistency Analysis

5.2 Comparison with Existing Methods

Table 3: Qualitative Comparison of Uncertainty Quantification Methods

Aspect	Entropy-Based	Semantic Consistency	Our Method
Dimensionality	1 (scalar)	1 (scalar)	**3 (vector)**
Paraconsistency	No	No	**Yes**
Incomplete Info	No	No	**Yes**
Model-Agnostic	Yes	Yes	Yes
Computational Cost	Low	Medium	Medium

Table 3: Qualitative comparison with existing UQ methods. Quantitative comparison with implementations of these methods on shared benchmarks is needed for definitive conclusions.

5.3 Limitations and Future Work

We acknowledge several important limitations of the current work:

1. **Limited Empirical Evaluation:** The evaluation is based on only six hand-selected questions. A comprehensive evaluation on standardized benchmarks (e.g., TruthfulQA, HaluEval) with hundreds of examples is essential before drawing general conclusions. Future work must include quantitative comparisons against strong baselines such as semantic entropy (Kuhn et al., 2023).

2. **Heuristic Component Definitions:** The formulas for computing I (particularly the max operator and 0.5 scaling factor) and the use of Silhouette Score for C represent heuristic choices that were derived empirically rather than from first principles. A

more rigorous theoretical derivation connecting neutrosophic logic axioms to these computational definitions would strengthen the framework.

3. **Embedding Space Assumption:** The framework assumes that cosine similarity in the embedding space is a valid proxy for semantic equivalence. This assumption may fail for subtle factual errors within a coherent cluster or for responses that are superficially similar but semantically distinct.

4. **Parameter Sensitivity:** Results may vary with different values of n (number of samples), τ (temperature), and θ (clustering threshold). A systematic ablation study examining the sensitivity of (T, I, F) scores to these parameters is needed.

5. **Computational Cost:** The framework requires n API calls per question (n=10 in our experiments), increasing latency and cost for real-time applications. Analysis of the latency/cost trade-off as a function of n would be valuable for practitioners.

6. **Threshold Sensitivity:** The paraconsistency detection (Sum > 1) depends on the clustering threshold θ. Different thresholds may yield different determinations of paraconsistency status.

6. Conclusions

We have presented a framework for uncertainty quantification in Large Language Models based on Paraconsistent Neutrosophic Logic. Our contributions include:

1. A three-dimensional uncertainty decomposition (T, I, F) that aims to capture consensus, contradiction, and ambiguity simultaneously.

2. Support for paraconsistency, enabling representation of coexisting contradictory evidence.

3. A model-agnostic implementation applicable to any LLM through standard API access.

4. Preliminary evaluation on six question categories suggesting the framework can differentiate uncertainty patterns.

Our exploratory results suggest that paraconsistency may be a common phenomenon in LLM responses (observed in 5/6 cases), but we emphasize that this finding is preliminary and requires rigorous validation on larger datasets and standardized benchmarks. Future work will focus on comprehensive empirical evaluation, theoretical justification of component formulas, and parameter sensitivity analysis.

References

Carnielli, W. A., & Marcos, J. (2002). A taxonomy of C-systems. In Paraconsistency: The Logical Way to the Inconsistent (pp. 1-94). Marcel Dekker.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of ICML (pp. 1321-1330).

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., & Chen, Q. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2309.01219.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for language model evaluation. In Proceedings of ICLR.

Priest, G. (2002). Paraconsistency and dialetheism. In Handbook of the History of Logic (Vol. 5, pp. 129-204). Elsevier.

Reimers, N., & Gupta, U. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP (pp. 3982-3992).

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Shelmanov, A., Panov, M., Vashurin, R., Vazhentsev, A., Fadeeva, E., & Baldwin, T. (2025). Uncertainty quantification for large language models. Tutorial presented at ACL 2025.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., & Majumdar, A. (2025). A survey on uncertainty quantification of large language models. arXiv preprint.

Smarandache, F. (2002). A unifying field in logics: Neutrosophic logic. In Philosophy (pp. 1-141). American Research Press.

Smarandache, F. (2013). n-valued refined neutrosophic logic and its applications to physics. Progress in Physics, 4, 143-146.

Xiong, Y., Liu, Z., Shi, A., Zhang, S., Cui, Y., Liang, S., ... & Huang, M. (2023). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv preprint arXiv:2306.13063.

Data Availability: All code, data, and supplementary materials are available at: https://github.com/mleyva/neutrosophic-uq-llm

Acknowledgments: We thank the reviewers for their constructive feedback. We acknowledge the use of OpenAI's API for response generation.

Conflict of Interest: The authors declare no conflict of interest.